UNIVERSITY OF AGDER

Identifying Geographic Terms within Natural Language Text

by

Ole-Alexander Moy

Master Thesis in

Information and Communication Technology

University of Agder

Grimstad, May 26th, 2008

Abstract

The huge amount of textual data available in digital form in today's world increases the need for methods that facilitate ease of access and navigability. Automatic extraction of keywords from text bodies is one promising approach. However, the relevance of keywords are context dependent, and extracting relevant keywords often requires a semantic analysis, simply because words may have different meanings in different contexts. It is well-known that resolving such word sense ambiguity automatically can be very challenging. When the topic of interest is geographic information, important keywords would be geographic terms like countries, cities, counties and states.

This thesis presents a probabilistic method for automatic identification of geographic terms within natural language text. The method uses a database of geographic terms to identify possible geographic entities. In contrast to state of the art, we resolve semantic ambiguity by using a Bayesian classifier that takes the context of ambiguous words into account. In our empirical results, we report a geographic term identification accuracy of 90%. We thus believe that the approach we present can be of importance for those working within the field of text analysis and data-mining, when accurate geographic term identification is of importance.

Preface

This thesis is submitted in fulfillment of the of the requirements of the degree of Master of Science in Information and Communication Technology at the University of Agder, Faculty of Engineering and Science, Grimstad, Norway. The project is supported by Integrasco A/S. Integrasco A/S has provided data material and supporting frameworks which were used to carry out various parts of the study. Supervisor on the project has been Ole-Christoffer Granmo at the University of Agder.

I would like to give a great thank you to Ole-Christoffer Granmo for excellent supervision and guidance throughout the project period. Input and expertise provided by Dr. Granmo has been invaluable. I would also like to thank Jaran Nilsen (Integrasco A/S) and Aleksander M. Stensby (Integrasco A/S) for valuable feedback during the project period.

Grimstad, May 26th, 2008

Ole-Alexander Moy

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the society we live in, the amount of textual data available in digital form is larger and growing faster than ever before. The information can be found via the Internet which consists of an overwhelming amount of web pages. These web pages contain information about everything from old kingdoms to newly developed scientific methods or unconfirmed rumors about a famous person or new products. The Internet, often referred as the web, is an unstructured and unorganized entity and the need for structure and organization has shown itself . There are several companies trying to use the data available via the web in several ways, but they all have the same problem. There is an overwhelming amount of textual data available in digital form today. It is therefore not possible with human supervision, resulting in the need for crucial and accurate methods for automated text analysis.

There are several companies specializing in so-called Word of Mouth (WoM), analyzing what people say about products. WoM is found in textual content around the web, especially in so-called blogs and online discussion boards. People from all over the globe are discussing every thinkable subject and this chat has in many ways become a new marketing channel. Consumers that have bought various products ask for advice from other consumers that either have bought the product, or have an opinion about the product in question. This information is crucial because more and more people are basing their decisions on reviews and

experiences shared by other people from all over the world. Word of Mouth has become a new marketing channel over the past years, where people search the web via their favorite search engine and read up on what others have said about the product they are considering to purchase.

Information provided by users in various geographic locations can play a valuable role for some of those analyzing online discussion boards. Various geographic locations have different cultures, climates and other geographic parameters that can be used to distinguish the users from one and other. An example of this can be various products where looks play an important role and taste changes between various cultures. In [8], a method was proposed and developed for placing discussion board users in their respective geographic location. The method used information gathered or mined from online discussion boards including geographic terms mentioned by the users. Geographic terms mentioned by the discussion board users yielded poor results because geographic term identification is a difficult and resource consuming process. The difficulties encountered in [8] lead to the definition of this thesis. Hopefully this thesis will also be a contribution to this exact area of application.

The task of extracting and identifying words as used either in a geographic or non-geographic manner is trivial for a human. Computationally however, this task is not as straight forward. This means that a basic word matching process against a list of known geographic terms is not sufficient. Such a method will return matches for terms not used as geographic terms in the context in which they are used.

The name of a geographic location, the geographic term, can be used in several ways, either as a geographic term or as a word. It all depends on the context in which it is used. Take for example common English words such as ”police” and ”going”, that can be found listed in many geographic gazetteers because they are also geographic locations. Only looking at words written with capital letter can seem like a viable solution for excluding several of the terms such as those mentioned, but proper nouns can also be ambiguous. Take for example ”Paris”. It meats the criteria by having capital letter, but can be either a name of a person or at least one geographic location. Geographic terms are not always used as

such in their given context even when they are not ambiguous. There are for example numerous hotels and news papers that use their location or origin in their name such as ”Radisson SAS Oslo” and ”The New York Times”. Because the geographic term is a part of a longer name, it is not used as a geographic term and should not be suggested as such.

To be able to extract geographic terms, such as countries cities counties and states, used within natural language text provides useful data. This data can for example be used in other classifiers or methods in which geographic terms make up some of the input. It is therefore an interesting and valuable method that must be developed.

1.1 Goal and Contribution

The goal for this thesis is to develop and present an efficient method for identification of geographic terms within natural language text. This means taking into account geographic names with multiple terms in their names and account for unformatted text. Unformatted text means text where geographic names can be written without capital letter. A sub-goal for this thesis is to limit the creation and use of various word lists for inclusion and exclusion of geographic terms.

The contribution from this thesis into the field of natural language text processing and pattern classification will be a method which uses previously developed text processing and pattern classification methods. The method presented is able to solve and hence identify geographic terms within natural language text without heavily relying on word lists.

1.2 Previous Work

There exists at least two papers directly related to the identification of geographic terms within natural language text, but neither propose a solution for the task solved in a probabilistic way. One method bases itself on lists containing terms

commonly found succeeding and preceding proper nouns called qualifiers. These qualifiers are used to exclude certain terms [6] that are non-geographic. The other method found in [9] also bases itself on lists, but these lists contain the exact names and terms to ignore in the geographic term extraction process. This means that geographic names that are part of a registered name are ignored together with the term list. Such a list must be created for each language they want to do geographic name extraction for. In addition, [9] propose a method that would work in the opposite way of the previously mentioned list used to exclude terms. The list contained terms commonly associated with geographic terms. This was found to be of no value for the overall success rate of the proposed solution for geographic term identification. There are also several Named-Entity Recognition and Classification(NERC) where proposed solutions try to recognize several sort of named-entities, such as geographic locations. These are however not specialized towards the identification of geographic terms and do therefore not differ between i.e. ”New York” and ”New York Times” or rely heavily on lists containing names to do so [7].

1.3 Target Audience

The target audience for this report is anyone working with text analysis in which geographic locations are amongst the desired information to extract from texts. The thesis is of a technical manner and experience with computer programming is recommended, however not a requirement. The problem, results for the solution and conclusion presented in this thesis should not require any experience with computer programing. The solution and results should be interesting for anyone working with textual analysis using contextual data.

1.4 Report Outline

The rest of this thesis is organized as follows; chapter 2 contains the background information required. Chapter 3 presents the challenges of geographic term iden-

tification within natural language. Chapter 4 is the solution chapter where the solution is proposed for solving the problem at hand is presented. The Prototype is presented in chapter 5. The results from the four examples presented in the prototype chapter is discussed in chapter 6 before the conclusion and further work is presented in chapter 7.

Chapter 2

Background

2.1 Word Sense Disambiguation

Word sense disambiguation (WSD) is a field within computational linguistics with the goal of identifying for which sense a word is used within the given context. For a human, this is normally pretty obvious given the in which context the ambiguous word is used what the meaning is. In [5] it is stated that Sense disambiguation is an *intermediate task* and in itself not an end, but necessary at some point for accomplishing most natural language processing tasks. Such tasks can be grammatical analysis and general text processing e.g. for spelling correction like case changes (”I AM IN LONDON” → ”I am in London”).

Like all other natural language processing methods, WSD can be classified into two main approaches, namely deep approaches (DA) and shallow approaches (SA). In essence, DA tries to understand the text at hand, which has proven to be a very difficult and an unsuccessful method computationally. SA on the other hand does not try to understand the given text, but looks at the composition of the words in the given text or the words surrounding a targeted term within a text.

The sentence ”The box was in the pen” is an example given by Bar-Hillel (1960) thought to be impossible to fully solve computationally for the word ”pen”, which can refer to the tool used for writing or an enclosure for animals, within the

given context and serves as a good example for word sense disambiguation and the problem of ambiguity.

2.2 Geographic Gazetteer

The GeoNames database of geographical names is available free of charge under a Creative Commons license [3]. The database can either be downloaded for offline usage or used via a web service. The database contains some eight million geographical names with populated places and alternate names for some of the geographical locations. Over 2 million of these names are classified as populated places with coordinates and population numbers. The GeoNames gazetteer consists names mined from WikiPedia which again is edited and maintained by its users on a volunteer basis. Thus, the information in the gazetteer may be incorrect and far from incomplete even with over 1.5 million distinct names classified as populated places. The GeoNames gazetteer was not designed for geographic name extraction, but for searching and have multiple instances of each name, one for each registered location.

2.3 Pattern Classification

The task of recognizing patterns and placing them into respective groups is called Pattern classification. In [12] Pattern Classification is described as a fundamental building block within machine learning and data mining. Despite being a complex problem, humans do this subconscious, constantly classifying, i.e. by recognizing a car as a car given all the shapes, sizes and colors a car can have. The different shapes, sizes and colors are called features and a car can have a variety of features either unique to the element or shared with others.

Pattern classification is divided into several areas with several valid approaches within each area. All areas do, despite different approaches, share some core elements for the Pattern Recognition and Classification process as shown in fig-

ure 2.1. The core elements possessed by both systems are; feature extraction, training and classification.

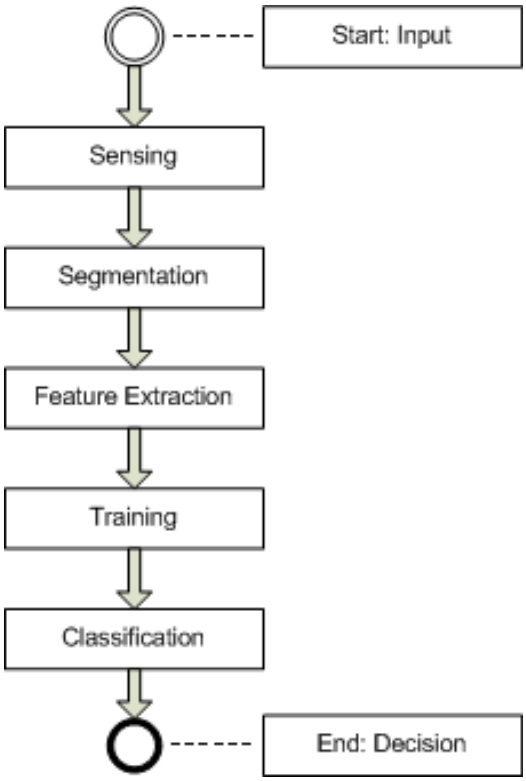

Figure 2.1: Common flow for pattern classification

Feature extraction for an object can be done in different ways depending upon how one can measure features for the object at hand. Using the case presented in this thesis as an example, surrounding words are the features that can be used in later classification or grouping. An example for where we are to do feature extraction is the sentence ”Following the end of the war in 1783, Washington retired to his plantation on Mount Vernon.” where the features are the terms surrounding ”Washington”. The features extracted can then be used by the classifier for placing input data within a specific group or class. There are several methods devised based upon the different decision theories such as Decision Trees and Bayesian Analysis.

By using the feature vector, a set of features characteristic for the input, provided by the feature extractor, the classification process manages to classify an

object to the correct category. A probabilistic approach such as naive Bayes' theorem can be used for the decision making process.

Training of the classifier can be done using half the provided data-set as training data and the other half for testing. This is the most common way for training and testing of pattern classifiers. On small data-sets, one can train the classifier on the whole corpus minus N entries selected, making sure that no two rounds contain the same collection of selected points, leaving us with $\binom{S}{N}$, where S is the corpus size, possible selections for testing. We will go deeper into bayesian theory later in this chapter.

2.4 Bayesian Theory

Bayesian classifiers are one of the two most used approaches used for solving word sense disambiguations together with decision trees. The approach is based on probability theory and the associated costs that relate to the different decisions. The Bayesian decision theory is based on known probability values called priori probabilities that are calculated before the decision process begins. In this section we present the background for our classifier.

2.4.1 Naive Bayes Classifier

The Bayes formula, also referred to as *Bayes' rule* and *Bayes' theorem*, is used within Bayesian decision theory for calculating probability [2].

$$P(\omega_j|X) = \frac{p(X|\omega_j)P(\omega_j)}{p(X)} \tag{2.1}$$

The equation above is used to calculate the *posterior* probability when the prior probability $P(\omega_j)$ and the conditional density $P(x|\omega_j)$ are known. Most significant for the equation is the product of the likelihood and the prior probability for determining the posterior probability; the *evidence* factor, $p(X)$, is not as sig-

nificant because it can be viewed merely as a scale factor guaranteeing that the posterior probabilities sum to one.

For the example, we assume that there are only two valid classes ω_1 and ω_2. X can be classified as ω_1 if and only if

$$f_b(E) = \frac{p(\omega = geoterm|X)}{p(\omega =!geoterm|X)} \geq 1 \tag{2.2}$$

The naive Bayes classifier is a simple probabilistic classifier based upon the application of Bayes' theorem with strong (naive) independence assumptions. For practical applications, parameter estimation for naive Bayes models uses a method named maximum likelihood. In this thesis, parameters refer to words surrounding the word we want to classify. This can leave us with a naive Bayes model without using any Bayesian methods. This can be done if we assume that all attributes are independent given the value of the class variable; that is;

$$p(X|\omega_j) = p(x_1, x_2, \cdots, x_n|c) = \prod_{i=1}^{n} p(x_i|c) \tag{2.3}$$

From this, can derive the resulting function $f_{nb}(E)$;

$$f_{nb}(E) = \frac{p(\omega = geoterm)}{p(\omega =!geoterm} \prod_{i=1}^{n} \frac{p(x_i|C = geoterm)}{p(x_i|C =!geoterm)} \tag{2.4}$$

The function $f_{nb}(E)$ is the naive Bayesian classifier, often referred to as naive Bayes. Naive Bayes is the simplest form of Bayesian network, in which all of the attributes are independent given the value of the class variable.

2.4.2 Bayesian Belief Network

A Bayesian belief network or Bayesian network is a probabilistic model were statistical dependencies between indicators efficiently can be represented and investigated [2]. The Bayesian networks simplifies the task of seeing how different

probabilities for different parameters affect the outcome of a probability model. This again is why Baysian belief networks are used within the field of pattern classification to see how the classifier reacts to various combinations of features selected for use in a classification process [4]. Probability theory provides an excellent basis for handling both randomness and uncertainty for various models, in our case the model found in figure 2.2.

The network structure is built by connection the casually related variables in a graph. An example based upon our initial thoughts of the task at hand represented as a Bayesian network can be observed in Figure 2.2. In figure 2.2 GT(Geographic Term) represents the final classifier giving the final decision, the end result, made by the classifier. The classification decision GT depends upon its first child nodes - g_1,g_2 and g_3 - which again represents a probabilistic decision made based upon their child nodes taking some parameters.

Figure 2.2. displays how various variables impact the leaf-nodes making initial decisions that again affect their parent node and finally combine into GT making the final decision and hence the outcome for the network of bayesian probabilities. By visualizing the graph for a Bayesian belief network it is possible to predict the outcome of the classifier GT. It can also be used for investigating which indicators were affect in a correct or erroneous way by the selected parameters.

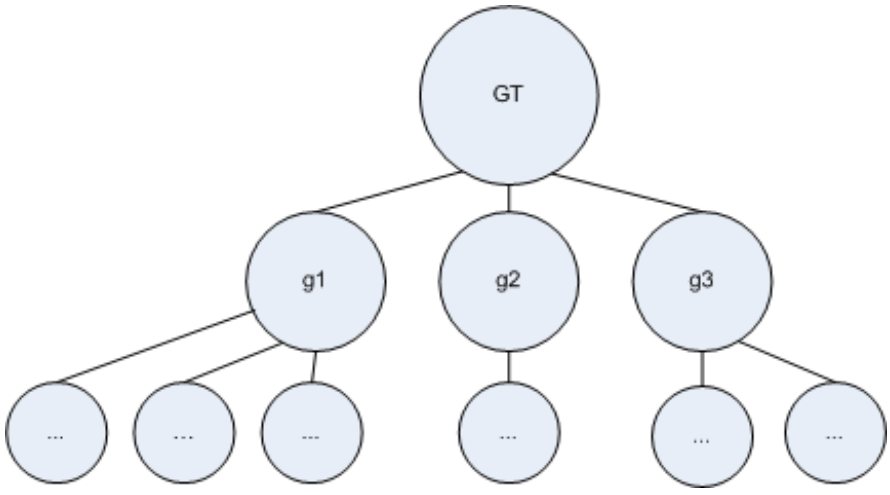

Figure 2.2: Bayesian belief network

In this thesis the Bayesian network is used in order to look at and investigate how different parameters and results from different classifiers combine into the geographic classifier GT shown as the root node in the figure used in the example above. The different dependencies between indicators used within each of the classifiers combined into the root classifier could not be observed in the same manner without the bayesian belief network.

2.4.3 Confusion Matrix

The confusion matrix, or table of confusion, represented in table 2.4.3 can contain information about actual or predicted classifications made by a classification system [11]. With actual data in the confusion matrix, we can observe how well a classifier is able to classify patterns - it's accuracy. The table can double not only as means of measuring accuracy, but the data it contains can also be used in cases where the classifier is indecisive or, simply put, confused.

Table 2.4.3 is used as a basic demonstration of a confusion matrix. The column named "A" is the true category for the classification pattern. Row A or the topmost row represents the actual classification results from the classifier. High values in the matrix' diagonal (cell a, cell b) indicates that the classifier has a high accuracy. The opposite conclusion can be drawn if high values can be found in either cell c or d. Results from the different cells can be used for weighting decisions by the classifier as mentioned earlier.

	A	**B**
A	a	c
B	d	b

Table 2.1: Example confusion matrix

2.4.4 Classification Error Rate

A classifiers classification error rate is a measurement that can be use to rate the accuracy of the classifier. The classification error rate is the percentage of new

patterns that have been assigned to the wrong category [2]. The goal for all classifiers is to be as exact as possible, hence striving for the lowest classification error rate.

Chapter 3

Geographic Term Identification

In this chapter we present the challenges surrounding identification of geographic terms and our proposed method. The data available for the study consists of several structured and semi-structured natural language text containing geographic terms. Data includes texts of various quality that can be from news paper articles to text taken from the Integrasco database.

A geographic term, or geographic name, is the name for a geographic location. A geographic term can consist of several terms like ”new” and ”york” which combines to ”New York”. Several geographic terms are also common terms in several languages, including English. There are several methods that seem obvious for solving the challenge at hand. A method that at first glance can seem viable is to match each term in the text with the previously mentioned list of geographic terms. In another method, proposed in [10], only capital words are suggested used and hence matched against a geographic gazetteer. Both proposed methods have flaws in that they will both give positive matches on to many words that in their given context are not used as geographic terms. Proper nouns should be written with capital letter and many share names with geographic locations. Many common words are also shared between their category or categories and geographic locations. Words such as ”going”, ”send” and ”police” are all common words in the English language and have at least one entry in a list of geographic terms, referred to as a gazetteer. In the following section we go deeper into the challenges

of geographic term identification.

3.1 Term Identification Challenges

In the list below we present and go deeper into the different challenges we have observed for the geographic term identification process. Proposed solution to the various challenges presented in this section is further explained in chapter 4.

Verbs

Several verbs double as geographic terms. They are therefore found in our geographic term list and will yield matches from the list of geographic terms. Some examples of verbs being geographic terms and one of their locations can be observed in table 3.1 below.

Name	Country
Going	Austria
Send	England
Bath	England
Run	Netherlands
Walk	Belgium

Table 3.1: Verbs as geographic terms

They therefore have the potential of being identified as geographic terms when they are not used accordingly and must be taken in to consideration when developing the classifier.

Nouns

Proper and common nouns are also used for geographic or non-geographic names in several cases. Many geographic names share their name with one or more non-geographic related name such as the proper noun ”Washington” which is the name of several geographic locations and the last name of the 1st President of

the United States. Several nouns include geographic terms within them, such as hotels and news papers that are named with the name of the city in which it is located i.e. ”Radisson SAS Oslo” and ”The New York Times”. Many geographic terms and common nouns are also ambiguous, for example ”Bath” in the United Kingdom and ”Cork” in Ireland, but the examples are numerous. Table 3.2 lists some famous people where both the first and last name yielded at least one hit from the geographic gazetteer.

Name	**City, Country**
Henry Ford	Henry, Haiti Ford, Ireland
George Lucas	George, South Africa Lucas, Bolivia
Lance Armstrong	Lance, Mozambique Armstrong, Argentina
Gordon Brown	Gordon, USA Brown, USA
Javier Solana	Javier, Spain Solana, Philippines

Table 3.2: Famous people with ambiguous names

Adjectives

In table 3.3 five adjectives are represented together with one of the geographic locations it is ambiguous with. The examples are only listed to give a small overview over how many common adjectives that actually have one or more geographic locations associate with them.

Name	**Country**
Hard	Austria
Soft	Zimbabwe
Nice	France
Blue	USA
Long	France

Table 3.3: Adjectives as geographic terms

The method presented in the next chapter will hopefully be able to differ between when the various adjectives and other terms are used as geographic terms and when they are not.

Prepositions

Prepositions are no exception from having entries in the list of geographic names and therefore can demonstrate . One of the prepositions used in the previous sentence; ”from” which in our gazetteer has one entry for a location within the city of Oslo, Norway. Not many of the prepositions found in the English language are ambiguous with geographic terms, but this can be very different for other languages.

Abbreviations

Several abbreviation are ambiguous with geographic names around the world. ”LOL” for example is a common abbreviation used on the Internet as slang for the English expression ”Laugh Out Loud”, but ”Lol” also yields a hit in the gazetteer with reference to a town in France. This abbreviation is used within many languages and is language independent even tho it refers to an English term. Abbreviations that are language dependent cause special ambiguous cases for various languages. This is not necessarily the case where some are used in several or all languages where they occur as with the example given above.

Language Dependencies

Many geographic terms are written different in several languages. Some of the geographic terms are used as they are in most or all languages as for example with the city of ”Los Angeles”, California, USA which is a Spanish name. The name is language independent, not translated or written in any other way in languages using the latin alphabet. Many country names are language dependent and are written in a completely different way in individual languages. Norway is for ex-

ample referred to as ”Norge” in Norwegian, ”Norvège” in French and Norwegen in German. This is also applicable, but not as common, for several towns and cities around the globe. An example of this is the city ”Munich” which is written ”München” in German.

Aliases

Geographic locations around the world are often referred to as just a part of their full name as for example ”Hull” which has the formal name ”Kingston upon Hull”. Another example is ”New York City” which often is referred to with the same name as the state in which it is located, the state of New York. Geographic terms existing of two or three terms are often abbreviated to the first letter in each of the terms. If the abbreviation is not registered in the gazetteer as an alternate name, we have no means of identifying the reference. Examples for this are ”Los Angeles” which often is referred to as ”LA” or ”New York City” which often is referred to ”NYC” or simply ”NY” as previously described. We will not focus on the aliases other than those provided for geographic terms in our gazetteer, though these mainly are linguistic aliases and not abbreviations used as aliases for geographic terms.

Descriptive Terms

Descriptive terms such as ”central”, ”outer” and the cardinal points can also double as place names on their own and are seldom registered in geographic gazetteers because they are vague descriptive names for larger areas within a geographic location. ”Central London” does not have an entry in the list of known geographic locations, but both ”central” and ”London” separately can be found in the gazetteer. ”Central” and other common words used in close range of geographic terms often yield hits in geographic gazetteers which can result in the descriptive terms being classified as geographic terms because they are used within the very same context as the geographic term they are describing as can be observed in the following sentence; ”A bottle-nosed whale is traveling up the River Thames in

central London, watched by riverside crowds.”.

Shared Context

To use a naive Bayes classifier we look at the words surrounding the term which we wish to identify. Many terms such as descriptive terms also double as geographic names, as explained above, and will therefore share many of the same contextual terms as the geographic term they are in front of have. An example of this can be observed in the example sentence in the previous section where ”central” and ”London” share the same context with an offset of one word. In front of both terms there are words which may have higher probability for being in front of a geographic or non-geographic term shifting the probability significantly.

3.2 Geographic Term Extraction Challenges

Geographic Term Extraction

The simplest term extraction process for geographic terms seems to be just checking every word in a sentence and see if it exists in the a list of geographic terms. This means that all geographic terms consisting of more than one term would be left out or that only parts of the geographic term would be extracted. If the basic match process would be used on ”New York”, ”New” would yield no results, but ”York” would. The proposed solution for geographic term extraction is presented in section 4.2

Tokenization

Tokenization is the very first part of parsing done when working with textual information. Demarcating and possibly classifying sections of the string into words, numeric expressions and punctuations leaves an output with valuable information about the sentence before further parsing.

The basic tokenization splits text text into sentences and then on white-spaces. ”The box was in the pen” is the sentence used in the following example given in an XML structure:

Listing 3.1: ”Tokenized sentence”

```
<sentence>
        <token>The</token>
        <token>box</token>
        <token>was</token>
        <token>in </token>
        <token>the </token>
        <token>pen</token>
</sentence>
```

There are several common issues that has to solved for the tokenization process to output useful data. In [7], accentuated characters, ligature and hyphenation are mentioned as some of the common issues within tokenization that depending on later use might have to be taken into account for the tokenization process. In the list below, we present our list of focus within the tokenization process.

- **Sentence Boundaries Detection**
  Sentence Boundary Detection (SBD) can either be done before or after the tokenization and has some challenges that has to be taken into account. Normally, a sentence ends with a period, exclamation or question mark, but these can also occur in the sentence without ending it. SBD seems like a fairly obvious task to solve, below are several examples that prove the opposite.

  ”Yahoo!” is an example of a trademark that breaks simple implementations of SBD if not taken into account. ”3,14” and ”3.14” are two ways of writing the same number, language dependent, but the punctuation that some BSD detect as a sentence break is in fact useful information in its given context.

- **Hyphen**
  Hyphens can play an ambiguous role as a part of a hyphenated word, pronominal inversions(grouping words with different part-of-speech), split words

at the end of a sentence(purely aesthetic) and instead of a apposition markers(replacement for comma).

- **Apostrophe** The apostrophe is a language specific character used in languages such as English as a possessive marker, in French as a determiner from a word that starts with a vowel and in Norwegian it is hardly used at all. There are exceptions from the rules defined above, e.g. there is one single exception in French, ”ajourd’hui (today)”, that breaks the general rule above for the language. Names are not language dependent and can break the rules defined for a language if not taken into account like ”O’Connor” and other Irish names.

Chapter 4

Proposed Solution

This chapter contains the proposed solutions to the challenges of geographic term identification presented in section 3.1. Section 4.1 presents a solution to the problem of tokenization. In section 4.2, the proposed solution for geographic term extraction is presented before the geographic term identifier is presented in section 4.3

4.1 Tokenization

In this section we briefly explain and present parameters used in the tokenization process of the texts we have gathered for training and testing purposes.

Sentence Boundary Detection is a difficult challenge because the text we are to train and test on is not very well organized and not always properly formated. We use available methods provided with Java to do basic Sentence Boundary Detection and accept that we are not able to detect all errors made by this process.

After dividing the text into sentences, each of the sentences are tokenized with a common tokenizer. The tokenizer splits the sentence on a set of given characters leaving us with a manageable list of tokens without special characters.

4.2 Geographic Term Extraction

The proposed method for geographic term identification we present in the next section bases itself on applying an identification process on each of the possible geographic terms found in a text by comparing the words in the text against a list of geographic terms. As descried in the previous section, geographic names can consist of one or more terms as for example ”San Fransisco”, ”New York” and ”Los Angeles”. In this section we present our proposed solution for extracting geographic terms consisting of one or more words within a text.

The geographic term extraction is the most resource consuming process because each word in the sentence that is to be analyzed has to be checked with the geographic gazetteer and its over 1.5 million distinctly named populated places. For the name lookup procedure we use the longest coincidence name matching method. Longest coincidence name matching of geographic terms means that we use the longest possible match for a geographic name in the gazetteer. This means that we will find and differ between ”York” and ”New York” or any other geographic term made up by more than one term. To accomplish this means that we have to do $n + 2$ queries for each possible geographic term we identify where n equals the number of terms making up the geographic term before applying the identification process. The method is used before we apply our identification method which will solve possible conflicts that this method might result in. We will, by doing this, be able to extract geographic terms such as ”North Wales” and ”South Wales”, which again means that the information retrieved can be even more accurate than compared to just identifying ”Wales”. The longest coincidence name matching algorithm is further explained in the pseudo code in figure 4.1.

Listing 4.1: Longest coincidence name matching pseudo code

```
// All matches are put in a stack
geotermStack := []

// Continue until end of the sentence or no matches
while possitionWithinSentence < lengthOfSentence

        // Get the current token
```

```
        token := tokens[possitionWithinSentence]

        // Set current token as start of
        // a geoterm.
        possibleTerm = token
        tempPositionWithinSentence :=
                possitionWithinSentence

        // Loop over and find the longest geoterm possible
        while gazetteer.hasGeotermStartingWith(possibleTerm)

                //  Add the next token to be checked
                possibleTerm := possibleTerm + " "
                        + tokens[tempPositionWithinSentence]

                // Increase the current position by 1
                tempPositionWithinSentence :=
                        tempPositionWithinSentence + 1
        end

        // Check to see if the geoterm has an exact match
        if gazetteer.hasExactMatchForLongestMatch(possibleTerm)
                // A geographic term was found, increase
                //the position
                possitionWithinSentence :=
                        tempPositionWithinSentence - 1

                // Add the found term to a stack
                geotermStack.add(possibleTerm)
        else
                // Nothing found, move to the next term
                // in the sentence
                tempPositionWithinSentence :=
                        possitionWithinSentence + 1
        end
end
```

As a restriction for the lookup task and hence a limitation for which geographic locations we are to identify we will only look up and try to identify geographic

terms with a registered population of above 2500. By doing this we limit our gazetteer in effect to contain some 54000 distinctly named geographic places. The total number of distinctly named registered populated places without regard to the population number is some 1.5 million, a vast amount of place names. Though we are ignoring a number of named places because of the population restriction, it will benefit us significantly and is an acceptable loss compared to efficiency. By applying this restriction, the identification process will only take in to consideration the most common population place names. There are however sufficient variability of geographic and non-geographic terms and cases of ambiguity.

In the following section we describe how we use the geographic terms extracted from the method described in this section and classify them as geographic terms or non geographic terms in the given context they are were extracted from.

4.3 Geographic Term Identification

In section 4.3.1 we present the indicators and how each of the indicators parameters are extracted for later to be used in our classifier presented in section 4.3.2. The intention for the classifier is to solve the various ambiguous challenges that reside within the geographic term identification process described in section 3.1.

4.3.1 Indicators

The surrounding terms of term extracted by the method described in the previous section are the basis for the indicators we need to classify the term. The surrounding terms are combined as bayesian properties, meaning that no word is found in more than one indicator. The classifier is applied to all geographic terms returned from the term extraction process described in section 4.2 to classify if the extracted term is used as a geographic term in its given context. We have selected three indicators - g_1, g_2 and g_3 - in which we wish to see if we can classify extracted terms as geographic or not. The properties making up the three indicators are; the first pre term or leading term for the current term in which we wish to classify, five pre

terms not including the first pre term used in the first indicator. The last indicator consists of the five post terms after the current term we are to classify. Below we give a concrete example on how the indicators are extracted and display the three indicators with their respective properties for the example sentence. Indicator g_1 will most likely provide important, but not contextually related terms such as prepositions and the two later is designed to bring enough contextual data for the classifier to be able to distinguish geographic from non-geographic terms.

Following the end of the war in 1783, Washington retired to his plantation on Mount Vernon.

Figure 4.1: Example sentence

Figure 4.1 shows the example sentence where the previously returned term from the geographic term extraction process is ”Washington”. The first pre term is collected and used as one indicator, the five next pre terms are collected and combined into an indicator as with the five post terms. Figure 4.2 displays the three individual indicators which we have selected and their respective properties.

First Pre Term

Term
1783

Five Pre Terms

Term
end
of
the
war
in

Five Post Terms

Term
retired
to
his
plantation
on

Figure 4.2: Example sentence parameters

In the left most table in figure 4.2, we can observe the term ”1783”. The middle and right most tables show the 5 words in each direction from the term we want to classify. The word found in the first table is skipped for the five pre term extraction. Each of the terms in the tables above that have been observed during training of the classifier will have a probability for, or against being used in the context of a geographic term. The training data is the key to how successful the classifier can be in classification. Training data is the key for all classifiers and

can therefore not contain to much noise before the classifier is unable to do proper classifications.

By combining all the probabilities for and against the extracted term being a geographic term, we can say whether the term most likely is used as a geographic or non-geographic term within the given context. The combining of parameters, indicators and the classification itself is further explained in the next section.

4.3.2 Geographic Term Classifier

The previously described parameters making up the indicators and the indicators themselves are combined in the final step making up the bayesian network, described in the background chapter, making up the classifier. This is where we can observe the success or failure of the selected parameters and combined indicators. A high success rate for the classifier means that we have been successful in solving the previously presented challenges of ambiguity.

Indicator g_1 is often bound by prepositions and other descriptive terms as explained in section 3.1. This can be a weakness for the indicator, but all of the indicators can have various weaknesses. The two other indicators have a high variety of terms because of the number of properties. The only difference between the two is that the five pre terms does not contain the term found in indicator g_1, but the basics of moving five terms in their respective direction is the same. Skipping the term added to indicator g_1 was done so that no parameter was used within two indicators. Each indicator alone does not provide enough of the context surrounding the term for classification, but combined we have enough information for and against the term being used in a geographic or non geographic way. The examples ”the United Kingdom” and ”the hull” share the same first pre term, but the rest of the surrounding words making up the context are most likely different. The first example can contain contextual terms such as ”country” which may have a higher probability for surrounding a geographic term than not. The term ”country” can be observed in the context of ”United Kingdom” in the following sentence; ”England, despite being the largest country of the United Kingdom, has no devolved executive or legislature and is ruled and legislated for directly by the

UK government and parliament.". In the second example, "hull" refers to a ships hull where several of the terms surrounding the extracted term can have a probability supporting "hull" used as a geographic term. An example of this can be observed in the following sentence; "The shape of a ship hull is determined by many competing influences." where "hull" refers to a ships hull.

The combining and hence calculation, of the final probabilities used by the classifier can be explained with the following formula.

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|parameters_{pro}) \tag{4.1}$$

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|parameters_{con}) \tag{4.2}$$

In equation 4.1, X_1 to X_n represent parameters, words in the context of the word that is to be classified, and their given probabilities. The returned result is the product of all the observations - first pre term, first five pre terms and last five pre terms - combined into two numbers. One giving a combined probability for and one against the current term being used as a geographic term based upon the observations from the context in which it is used. To see which of the two end probabilities is the highest we can do $log(pro/con)$ and see if the end result is bigger, equal or lower than one. The next chapter presents validation cases for the four types of cases that can occur for the classifier.

Chapter 5

Prototype

A prototype has been developed as a proof-of-concept for the method presented in prior chapters. The objective of this prototype is to validate the initial assumptions.

Formulas and methods previously described in the thesis have been implemented as algorithms and methods using the Java programming language together with underlying Java frameworks. The basic structure of the prototype can be viewed in the model presented in figure 5.1 below. The proposed proof-of-concept implementation is designed in a way which hopefully will make it easy to incorporate into projects where indications of geographic terms is required.

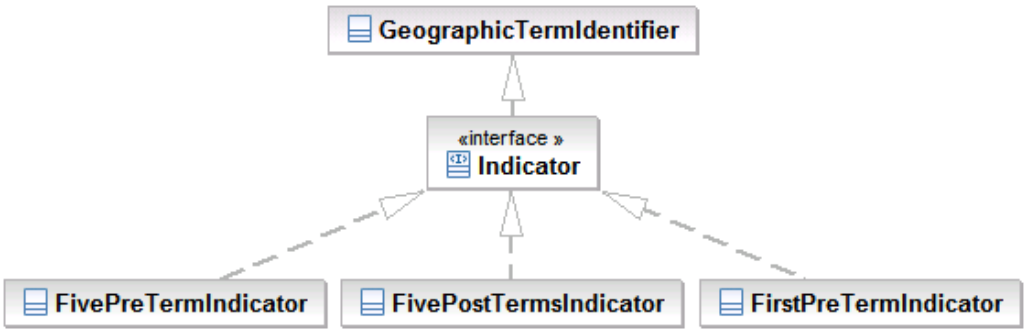

Figure 5.1: Overview model of prototype

In the following sections we will cover the testing that has been performed on

the prototype. We will come back to the different indicators and how they affect the classification process in the next chapter.

5.1 Testing and Validation

The basis for the analysis is the Bayesian belief network as previously described in section 4.2, which we have implemented in our proof-of-concept prototype.

The training corpus for the classifier consisted of 200 sentences containing one or more samples of geographic terms and 200 lines with one or more samples of ambiguous verbs, proper and common nouns and adjectives. The testing was done against a list of sentences contag 110 sample cases. A case is this context means returned hits from the geographic term extraction process. The samples given for each of the classes were samples representing each of the two classes and were not included in the training of the classifier.

Naive Bayes early proved to be a good approach combined in a Bayesian belief network explained in section 2.4.2. With the Bayesian belief network we can easily observe which of the parameters have the greatest or least effect on the decision made by the classifier on each of the identified geographic terms. This quickly proved to be valuable for observing which words had probabilities resulting in correct or erroneous classification. Each indicator is made up by one or more parameters which all are added together in the Bayesian belief network.

Listing 5.1: Combining the parameter pseudo code

```
// Initial values, can be anything but 0
pro = 1;
con = 1;

// Looping over each parameter in each indicator
// adding the parameters together
forech parameter in parameters

        // The combined probability for each parameter
        //  in this particular indicator
```

```
        pro = pro * parameter[0]
        con = con * parameter[1]
end
```

The above pseudo code above shows how each parameters probability is combined and gives the combined probability of all the parameters making up the indicator. In the figure below, the total probability for(pro) and against(con) a term identified within the gazetteer being a geographic term or not within the given context.

Listing 5.2: Combining the indicators pseudo code

```
// Initial probability for a term
// being geographic
pro = initialPro

// Initial probability for a term
// being non-geographic
con = initialConValue

// Looping over and getting probabilities
// from each of the indicators
foreach indicator in indicators

        // The combined probability for each indicators
        // consisting of parameters accounting for
        // initial probability
         pro := pro * indicator[0]
         con := con * indicator[1]
end
```

The two probabilities can be compared by doing $log(pro/con)$ and check to see if the result is above or below 1. A result below 1 means that the con probability was larger than pro probability and hence the term is classified as a non-geographic term. If the result equals 1, the combined probabilities were the same and no decision can be made by the classifier.

By using the geographic term extraction method presented in 4.2, we were able to have only the words actually found in the geographic gazetteer to check.

The geographic term extraction process is, as previously mentioned, the most cpu-intensive process and it is uncritical in which results it returns as long as it finds at least one instance of it. This can lead to a vast amount of hits and hence for the classifier to classify. In the next sections we present four individual cases in which the classifier can meat and validate these before we discuss them further in the next chapter. Validation case 1 represents the true positive case, a case where a geographic term has been classified as such. Case 2 represents cases where words used as non-geographic terms are classified as geographic terms. The third validation case represents cases where non-geographic terms are classified correctly as a non-geographic term. Validation case 4 presents the case where a geographic term is classified as non-geographic terms, as a false negative. The parameters making up the indicators for each case can be observed in the tables represented in the figures in the following cases. The classifier uses ”0,55” as the initial pro probability and ”0,45” as the con probability.

5.1.1 Validation Case 1 - True Positive

In the first validation case presented a case where the classifier correctly classifies an identified geographic term as actually used as a geographic term in its given context is presented. The geographic term in the case is ambiguous, but the same operation is applied to all identified geographic terms. The sentence we have analyzed in this case is; ”Since the oresundsbro was completed Malmo has become a more vibrant place because it is just across the strait to Copenhagen.” where Malmo is our possibly identified term for the properties and indicators displayed in figure 5.2.

First Pre Term

Term	Pro	Con
completed	0.5	0.5

Five Pre Terms

Term	Pro	Con
since	0.00260	0.00130
the	0.03855	0.02650
oresundbro	0.5	0.5
-	-	-
-	-	-

Five Post Terms

Term	Pro	Con
has	0.01285	0.11682
become	0.5	0.5
a	0.03146	0.04606
more	0.00357	0.00357
vibrant	0.5	0.5

Figure 5.2: Case 1

Figure 5.2 displays three tables which represents the three indicators that we

are using in our validation example. The parameters in each indicator can be combined for the pro and con combined probability for the target term which yielded at least one hit in the gazetteer. The indicators can again be combined giving the classifier a mean to classify the target term. The left most table shows the probability for first pre term, the middle and most right table shows the five pre terms when skipping the term from the first table and the five post terms surrounding the term we are trying to classify.

5.1.2 Validation Case 2 - False Positive

With the following example, we show how the context surrounding an ambiguous term can be used to classify it either as used in a geographic context or not. ”plantation” is in this case used as a proper noun as observed in the following sentence; ”Following the end of the war in 1783, Washington retired to his plantation on Mount Vernon.”. Our example geographic term yielded at least one hit from

First Pre Term

Term	Pro	Con
his	0.5	0.5

Five Pre Terms

Term	Pro	Con
in	0.06772	0.00451
1783	0.5	0.5
washington	0.5	0.5
retired	0.5	0.5
to	0.02832	0.04926

Five Post Terms

Term	Pro	Con
on	0.00475	0.02612
mount	0.5	0.5
vernon	0.5	0.5
-	-	-
-	-	-

Figure 5.3: Case 2

the geographic gazetteer. The total sum for and against that ”plantation” in this case is used as a geographic term is the combined probabilities from each of the indicators.

5.1.3 Validation Case 3 - True Negative

The third validation case is the term *Washington*. The three tables found in figure 5.4 shows the surrounding terms for the target term in the following sentence; ”Following the end of the war in 1783, Washington retired to his plantation on Mount Vernon.”.

First Pre Term

Term	Pro	Con
1783	0.5	0.5

Five Pre Terms

Term	Pro	Con
end	0.00130	0.00390
of	0.01395	0.01035
the	0.03855	0.02650
war	0.5	0.5
in	0.06772	0.00390

Five Post Terms

Term	Pro	Con
retired	0.00576	0.01267
to	0.02613	0.04090
his	0.00119	0.00119
plantation	0.5	0.5
on	0.00475	0.02612

Figure 5.4: Case 3

5.1.4 Validation Case 4 - False Negative

In this example sentence, we look at the sentence ”The only people it will ACTUALLY affect are those who live in London, not you lot in your leafy suburbs.” which is about the city of London in Great Britain. Terms that have never been observed before get the same pro and con probability as displayed in the tables below.

First Pre Term

Term	Pro	Con
in	0.22571	0.01428

Five Pre Terms

Term	Pro	Con
affect	0.5	0.5
are	0.01023	0.01278
those	0.00130	0.00130
who	0.00388	0.01278
live	0.01147	0.01147

Five Post Terms

Term	Pro	Con
not	0.00825	0.01415
you	0.00475	0.03325
lot	0.00119	0.00119
in	0.00461	0.00346
your	0.00119	0.01315

Figure 5.5: Case 4

Chapter 6

Results and Discussion

In this chapter we are to present and discuss results from testing done with the prototype presented in the previous chapter. The results are discussed and analyzed in order to conclude how the classifier.

6.1 Indicator Results

Here we present and comment on each of the indicators and their results. The training of the indicators were done with a corpus containing 200 sentences with known geographic terms and 200 sentences with examples of ambiguous terms where the pre and post terms were extracted and counted. The variance in number of terms registered for the five pre and post terms are due to varying length of the training sentences. The Oxford English Dictionary[1] contains over half a million words from across the English-speaking world which means that we have only observed only a small number of the possible terms that can be used within geographic and non-geographic terms available in the English language. This also means that training the classifier to know about all terms is not feasible. Instead we have focused on training on specific cases manually picked to prove the concept of using the naive bayes classifier for geographic term identification.

6.1.1 Indicator g_1

Indicator g_1 consists of only one term. Of the 192 distinct terms found in the 400 sentences of training data only 16 of the terms had a higher probability for a term being in front of a geographic term compared to in front of a non-geographic term. The 16 terms with a higher probability for being used in front of a geographic term where among others ”in”, ”of” and ”from”. 54 of the terms had a higher probability for being the first pre term in front of a non-geographic term. Many terms were observed only once or twice in front of a geographic and non-geographic term in the testing data, 122 were found to be indifferent for the classification, meaning that they had an equal probability for or against being used in front of a geographic term. The indicator is significant and hence aids the classifier in cases where the more contextual indicators explained later lack known terms or the contextual terms are vague. The indicator does not contain the probability for the same variety of terms and can in some cases extract terms that have never been observed before, rendering the indicator indecisive for the current case. As previously explained, several of the descriptive terms are found amongst the 16 terms with a higher probability for being used in front of a geographic term, such as ”over” and ”central”. The most common term found in front of geographic terms in the training set are ”in” due to many training examples about mobile phones in different markets, split into countries. In the list below we comment on the results from the four cases presented in the previous chapter for this indicator.

- **Case 1:**
  The first pre term in this case is ”completed” which was never observed during training and hence is given a probability of 0.5 for being used in front of a geographic and non-geographic term. This means that the results of the indicator in this specific case is unimportant for the classifiers final decision.

- **Case 2:**
  In case 2 presented in section 5.1.2,the first pre term have, as in case 1, never been observed before and thus does not provide any useful information for the classifier to classify on.

- **Case 3:**
  Case 3 also suffers from the same problem as the two previous examples, the first pre term have never been observed before in front of either a geographic or non-geographic term. The indicator therefore does not provide anything useful for the classifier in this case.

- **Case 4:**
  The fourth case is unlike the others and show an example with the term ”in” which have been observed before. From the table in figure 5.5 we can observe that ”in” has a higher probability for being the first pre term in front of a geographic term than for a non-geographic. Finding known terms is of course the best case for the classifier, but because of the vast amount of terms, the classifier will not be able to observe all known terms during training.

6.1.2 Indicator g_2

By looking at the five pre terms and using these as parameters for this indicator, the goal was to gather enough significant and contextually relevant terms for the classification process. In the 400 training sentences a total of 776 distinct terms were extracted in a position up to five positions in front of the target term. Only 76 have a greater probability for being one of the five pre terms for a geographic term compared to being used as one of the five pre terms for a non-geographic term. There were 140 terms with a greater probability of being used as one of the five pre terms for non-geographic terms compared to being used in front of a geographic term. The rest of the terms are indifferent for the classifier in the same way as the first pre term have indifferent terms in the term list. The list below we discuss the results for this indicator in each of the previously presented cases.

- **Case 1:**
  In first test case we observe that only three terms were extracted. Both ”since” and ”the” have been observed within five pre terms of a geographic term more often than for non-geographic terms and hence have a higher

probability for being used in front of a geographic term. The last term observed, ”oresundbro”, have never been observed as one of the five pre terms and is insignificant for the total probability. In total the indicator achieves a higher probability for the target term being a geographic term.

- **Case 2:**
  With this particular case, the classifier erroneously classifies the target term as a geographic term. Three of the five pre terms found have never been observed during the training of the classifier. The two other terms are ”in” and ”to”. The first have a much higher probability of being found as one of the five pre terms for a geographic term and the later have a lower probability. The combined probability for the indicator leaves us with a higher probability for this term being a geographic term. This results in the erroneous probability from this indicator.

- **Case 3:**
  Out of the five first pre terms, four have been observed during training. The term that was not observed during training gets the same probability for and against being used in front of a geographic term. Of the four other terms, three have a higher probability for being in front of a geographic term. By combining the indicators we can confirm that this indicator provides the classifier with an erroneous probability for this case.

- **Case 4:**
  Case 4 shows and example where the classifier classifies a geographic term as non-geographic term. Here too as with other example cases one of the terms have not been observed either for geographic or non-geographic terms during training of the classifier. The combined probability for the four observed terms leaves the indicator with a total probability of not being a combination found as the five pre terms, thus the indicator in this case is incorrect.

6.1.3 Indicator g_3

As for the previous indicator the total number of terms found after the target terms vary because of the number of terms in the sentence after the targeted term. During the training 834 distinct terms were observed within the five post terms. In the list of terms, 75 term had a greater probability for being one of the five post terms for a geographic terms while 176 where found more often as one of the the five post terms for non-geographic terms. The other 583 terms were found to be insignificant for the classifier because the probability for being after a geographic or non-geographic term are the same. As mentioned for indicator g_2, the English language consists of a large amount of words. There are over half a million words registered in the Oxford English Dictionary meaning that we have only observed a small fragment of the total number of terms. Because of this, several of the post terms observed for the four cases have were not observed during the training of the classifier.

- **Case 1:**
  The first case two of the five post terms were never observed during training and are hence assigned a probability of 0.5 both for being used after geographic and non-geographic terms. The classifier correctly classifies the target term as geographic, but not due to this indicator which has a combined probability higher for the term not being a geographic term.

- **Case 2:**
  The classifier in this case falsely classifies the target term as a geographic term. Here two of the terms have never been observed neither after a geographic or non-geographic term. Only three terms were extracted due to the length of the sentence after the target term, but the term that have previously been observed has a higher probability for being one of the five post terms for a non-geographic term. The probabilities combined for the indicator correctly suggest that the term in question is in fact a non-geographic term, but not by enough to shift the total probability.

- **Case 3:**
  Case 3 presents a case where the classifier correctly classifies a non-geographic term. Only one of the terms in this indicator had been not been observed during the training of the classifier. The last term have been observed an equal amount of times as one of the five post terms for both geographic and non-geographic terms and therefore does not bring anything special for the classifier to work with.

- **Case 4:**
  In the fourth case the classifier falsely classifies a geographic term as a non-geographic term. In this particular case, all of the terms extracted for the indicator have been observed before. The term ”lot” have been observed an equal amount of times in the context of geographic and non-geographic terms as with one of the terms in the previous case. Only the term ”in” supports the indicator in getting a higher probability for the target term being a geographic term. The three last terms all have a higher probability for being one of the five post terms for a non-geographic term. The classifier in the end falsely classifies the target term as a non-geographic term.

6.2 Combined Results

The results from the different indicators are combined as described in section 4.3.2 are presented here. The context can vary in both the number of known terms and the actual number of terms available in the sentence in which the classifier is to extract parameters for its indicators. Because of the large number of words in languages, the classifier can not be based upon observing all the terms. The classifier must therefore be trained towards knowing as many possible terms surrounding geographic and ambiguous words, words used as non-geographic terms that also are names for geographic locations. In section 3.1 we presented the various challenges that we were encountering and that ambiguous geographic terms could prove to be the most difficult to solve, especially in cases where the context lacked supporting terms for the classifier to use for correct classification. We have

taken into account surrounding words of extracted geographic terms and have tried to classify them.

It was commented earlier that indicator g_1 would contain several prepositions and a minimum of geographic and non-geographic contextual information. The prediction made for the two other indicators were that they would contain more contextual data such as for example ”travel” and ”leaving” for geographic terms and for example ”ship” and ”buy” for non-geographic terms. Another prediction presented was the fact that the classifier could have a difficult time classifying between geographic and non-geographic terms when used in the same sentence and especially in close proximity of each other.

In table 6.1 we can see the result of the classifier running on our test data where sentences only have hits representing one class. The confusion matrix contains data representing all of the classifications made by the classifier either correct or erroneous. The classifier correctly classifies over 90% of the training data for each category leaving us with some false positive and false negatives.

	Geo	**Non**
Geo	89,19	10,81
Non	8,22	91,78

Table 6.1: Combined indicator results

The combined accuracy for the classifier given the confusion matrix above is 90.53%, leaving us with a 9,47% error rate for the test set.

The success rate quickly diminishes where both classes are represented in the same sentence. This was observed for cases where terms of different classes are represented in close proximity of each other in the same sentence. An example of the problem that occurs can be observed in the following sentence; ”A seven-tonne whale has made its way up the Thames to central London, where it is being watched by riverside crowds.” where ”London” and ”central” share many of the contextual words in the indicators, leaving the total probability very much the same. This results in either ”London” or ”central” to be classified erroneously as the wrong class.

Chapter 7

Conclusion and Further Work

In this chapter we present the conclusion based upon the previously given background information, information presented in chapter and the results discussed in the previous chapter.

7.1 Conclusion

With this thesis we have developed and presented means for identification of geographic terms within natural language text. The method bases itself upon information provided by natural language text. The information gathered is combined using a probabilistic approach and analyzed in order to classify terms either as geographic or non-geographic within the context they are used. In the solution chapter we present methods for solving the identification of geographic terms by using three key indicators; indicator g_1, indicator g_2 and indicator g_3. These three indicators use parameters found in the sentence in which the term we wish to classify resides. The sub-goal of limiting the number of lists required has also been met.

As presented in section 1.2, we are aware of other methods where geographic term identification within natural language text has been solved. However we are not aware of any previous work where this have been solved using this kind of

approach. We therefore believe that the proposed solutions and results for and of this study will have significant value for the project employer, and also other parties which are looking for methods for identifying geographic term identification or other similar problems of classifying words in different contexts.

The results presented in this report for the classifier are over all very satisfying. There is one issue in which more work must be done, namely solving ambiguity of terms used within the same sentence. This has proven to be a challenging problem to solve because so much of the context is shared between the geographic and non-geographic terms that reside within close proximity of each other. The algorithm has proven to be efficient and accurate for cases of ambiguity not residing within the same sentence and close proximity, as previously mentioned. However the classifier will be useful for the project employer for further use because it removes a great amount of sentences with hits on non-geographic terms used in a non-geographic context.

With the indicators developed we observe the randomness and uncertainty for the parameters making up the various indicators. By combining the indicators we are able to cope the randomness and manage the uncertainty for geographic terms within natural language text.

7.2 Further Work

With the indicators developed during the project we have been able to classify geographic terms within natural language text, but there is, as stated in the previous section, still room for improvements both for the classifier as presented here and for the special case mentioned in the previous section where there are terms representing each of the two classes in the same sentence. In this section we recommend some further work on the presented solution and in the area.

Shared Context

The classifier is not able to correctly classify geographic and non-geographic terms when terms representing each of the two classes are found in close proximity of each other in a sentence. Geographic terms are classified as non-geographic and non-geographic terms are classified as geographic terms based on the probability of the majority of the terms in the sentence. A method for capital letter indication can aid the classifier in correctly classifying common nouns within these sentences. Another approach can be to ignore ambiguous words written without capital letter. This means ignoring a minor portion of the proper ambiguous names that erroneously was written without capital letter. This does however not solve the problem of ambiguity for proper nouns. This problem is still unexplored.

Improved Training Sets

A classifier can not work properly or with a high accuracy without a large and proper training set. For text classification this is very important because of the vast amount of possible terms that can be used in sentences. It is therefor important to have a large training set to improve the classification success rate further.

Multilingual Support

The classifier should, given training examples for each of the two - geographic and non-geographic - classes, be able to identify geographic terms within natural language text for most languages without any change to the prototype itself. Because the training data is language dependent, a language identifier is needed to ensure that the correct probabilities used by the classifier. An N-Gram [1] solution can be appropriate for this task. There are several ways of finding conflicting terms that can be used for the non-geographic term class training set. One solution to finding conflicting terms can be using a list of terms for the current language

[1] An N-Gram is a sub-sequence of n items from a given sequence.

such as WordNet [2] for English. Another solution to this could be creating a list of the most common terms for the language and match these against the gazetteer that is to be used.

Positioning of Identified Geographic Terms

A desired feature for the geographic term identifier is the ability to return in which country (depending on the locale layout for administrative regions for countries, municipalities, counties and states can also be used) the identified term most likely can be placed in. One can e.g. combine population numbers together with for example memory stack of the last five countries observed during text classification depending on the type of text as presented and used in [6]. Given the example of geographic term identification in a news paper article, the countries ”USA” and ”Norway” been identified as most likely being used as geographic terms. Later in the text ”Oslo” is observed, which in fact is registered geographic locations in both countries. Because Oslo, Norway has over half million citizens it’s most likely Oslo, Norway is being referred to, not Oslo, USA.

[2] WordNet is a large lexical database of English and can be found at http://wordnet.princeton.edu/ .

Bibliography

[1] Oxford english dictionary. [Online]. Available: http://www.oed.com/about/

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiely-Interscience, 2001.

[3] Geonames. GeoNames. [Online]. Available: http://www.geonames.org

[4] Z. Ghahramani, "Learning dynamic Bayesian networks," *Lecture Notes in Computer Science*, vol. 1387, pp. 168–197, 1998. [Online]. Available: http://citeseer.ist.psu.edu/article/ghahramani97learning.html

[5] N. Ide and J. Veronis. (1998) Word sense disambiguation: The state of the art. [Online]. Available: http://sites.univ-provence.fr/veronis/pdf/1998wsd.pdf

[6] Y. Kanada, "A method of geographical name extraction from japanese text for thematic geographical search," in *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management*. New York, NY, USA: ACM, 1999, pp. 46–54.

[7] D. Nadeau. (2005) Balie - baseline information extraction. [Online]. Available: http://balie.sourceforge.net/dnadeau05balie.pdf

[8] J. Nilsen, "Locating discussion board users with bayesian analysis of geographic terms, language and timestamps," p. 59, 2007. [Online]. Available: http://www.integrasco.no

[9] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A.-C. Forslund, and C. Best, "Geocoding multilingual texts: Recognition, disambiguation and visualisation," p. 53, 2006. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0609065

[10] B. Pouliquen, R. Steinberger, C. Ignat, and T. D. Groeve, "Geographical information recognition and visualization in texts written in various languages," in *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2004, pp. 1051–1058.

[11] A. Roberts. (2005) Guide to weka. [Online]. Available: http://www.andy-roberts.net/teaching/db32/weka-db32.pdf

[12] H. Zhang, "The optimality of naive bayes." [Online]. Available: http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf